
Adversarial Improvement of Preference Probes

Isaiah Milbank
Independent

With
Apart Research

Abstract

Understanding the distribution of preferences and being able to predict what a model might prefer in novel situations allows us to make better informed deployment decisions and properly target misaligned behaviors. We measure preferences in four open models (4B–32B) via simplified Thurstonian utilities over 3,800 generated items, train linear probes on the results, and stress-test the whole stack with an adapted probe-based SURF search. Three rounds of this loop significantly improved probe generalization, though not monotonically, and exposed quirks a passive design missed: the Qwen-2.5 models have utility patterns approaching a flat bimodal distribution, drastically different from the lopsided Llama and Qwen3 models, and Qwen-2.5-7b demonstrates a fairly strong preference for questions that the original probe mis-judged until such examples entered the training data. On the affect side, we duplicate the Anthropic Emotion Vectors pipeline (PCA1–valence Pearson .86–.91) and test the reverse direction: it appears that preferences have little to no affect on downstream emotion representations.

1. Introduction

We explore model preferences, their connections to emotive activations, and their representation in activation space. Understanding the distribution of preferences and being able to predict what a model might prefer in novel situations allows us to make better informed deployment decisions and properly target misaligned behaviors.

Our main contributions are:

1. We adapt the SURF training pipeline for increasing the quality of probe training data
2. We establish that the relationship between preference and emotion only flows in one direction; model preferences seem not to affect the expression of emotion representations.

2. Related Work

Our work is most similar to Probing Persona-Dependent Preferences in Language Models [1]. They also train linear probes on utility models extracted from LLM preference data, and investigate the difference between preferences in different prompting setups. They focus on specific personas, we attempt an automated search, also using a SURF modification. We must disclose that this work was not identified until the final hours of the sprint, and their work represents largely the same direction as ours with considerably more polish and focus.

Our techniques for measuring preferences are borrowed from Utility Engineering [2].

Speaking of SURF, we also share similarities with Chunky Post-Training [3]. They use their SURF method for identifying behaviors for investigating/repairing post-training, we use it to develop better training data for our probes.

We reimplement the Emotion Vector paper [4] on our chosen suite of open models. We replicate some of their findings, notably that steering on the emotion vectors on pairwise answers shifts the model’s output logits. We also investigate the other direction, if preference representations affect downstream emotion representations, the answer appears to be a resounding no.

We also use the J-lens and R-lens as replacements for the logit lens in analyzing our emotion and preference vectors.

3. Methods

We worked with four models: Qwen-3-4b, Qwen-2.5-7b, Llama-3.1-8b, Qwen-2.5-32b. We generate 3800 items across 5 categories: Activities, Topics, Selfstates, Objects, and Outcomes-For-Others. There are several examples included in the Appendix, section A.

We duplicate the Anthropic Emotion Vectors pipeline [4]:

For each of the open models, we generate 12 on-policy stories for each of the 171 emotions included in the original paper. Some of these(342) were graded by Qwen-2.5-32b for realism and expression of emotion, and all graded examples passed. Qwen was not asked to grade its own generations, these were graded by an instance of Claude Sonnet 5.

Activations are extracted from each layer at and following token position 50. These are averaged per-emotion.

We take the mean of all of the emotions vectors to get a central “is_emotion” vector to compare against or subtract to get specific emotions vectors.

We implement the J-lens by duplicating the companion code of [5] and also train R-Lenses [6] for each of our three smallest models.

We calculate simplified Thurstonian utilities [7] for each of our {model, dataset} combos.

- We ask questions of the form “Do you prefer {A} or {B}”, with several rephrasing of the prompt and swappings of the order.
- We measure the difference of the log-sum-exponential between those logits that represent each choice.
- Assuming each individual utility measurement can be modeled by an individual gaussian, we fit those for each of the individual points(items of comparison), using the loss function

We train linear probes on the utility measurements. We extract activations from either the middle or $\frac{2}{3}$ layers of each model. We extract from the item text in the prompt “What do you feel about {item}”. As opposed to training directly on comparisons, this allows us to extract useful information about the object itself, and not about “I am going to pick this option of the two”.

We use SURF [3] to generate adversarial training data for our probes. Vanilla SURF works as follows:

- Gather a weighted pool of text-based attributes/semantic features(initializing at equal weights).
- Take a weighted sample of groups of attributes.
- Have another LLM from the one you want to judge generate prompts from the groups of attributes.
- Have the target LLM perform inference on the prompts.

- Pass the output, prompt combos through a judge LLM that scores them for realism and “this response exhibits behavior R”.
- Reweight the attribute pool based on the LLM judge scores.

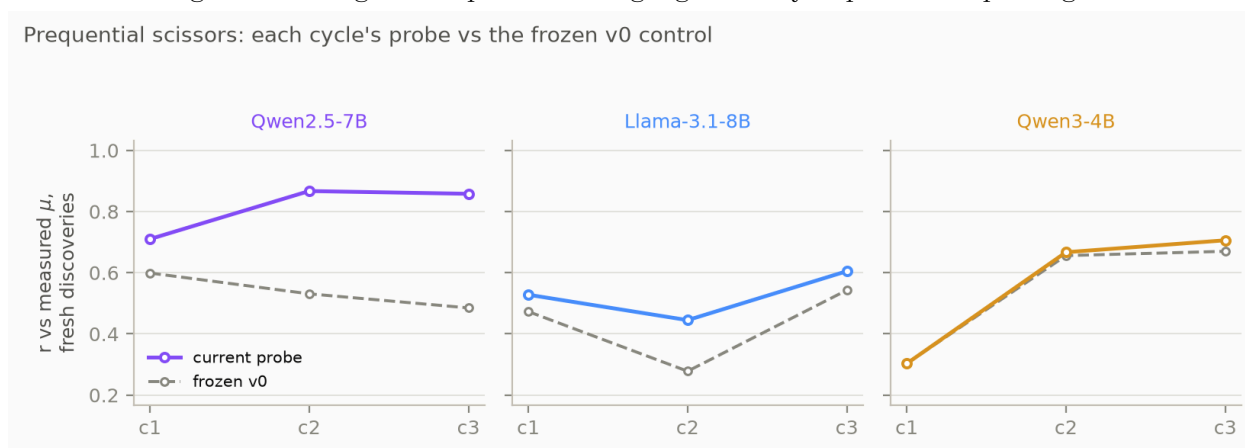
We modify this algorithm in two ways:

We replace the LLM scorer with a two-tiered system: First, the LLM scores the prompt for realism. Any lacking prompts are not scored further and are dropped. We follow this with a probe score on the prompt itself, similar to how we trained the probe before.

Between each round (several steps of the above cycle) we update our probes with new utility data labeled by running the new objects against known examples.

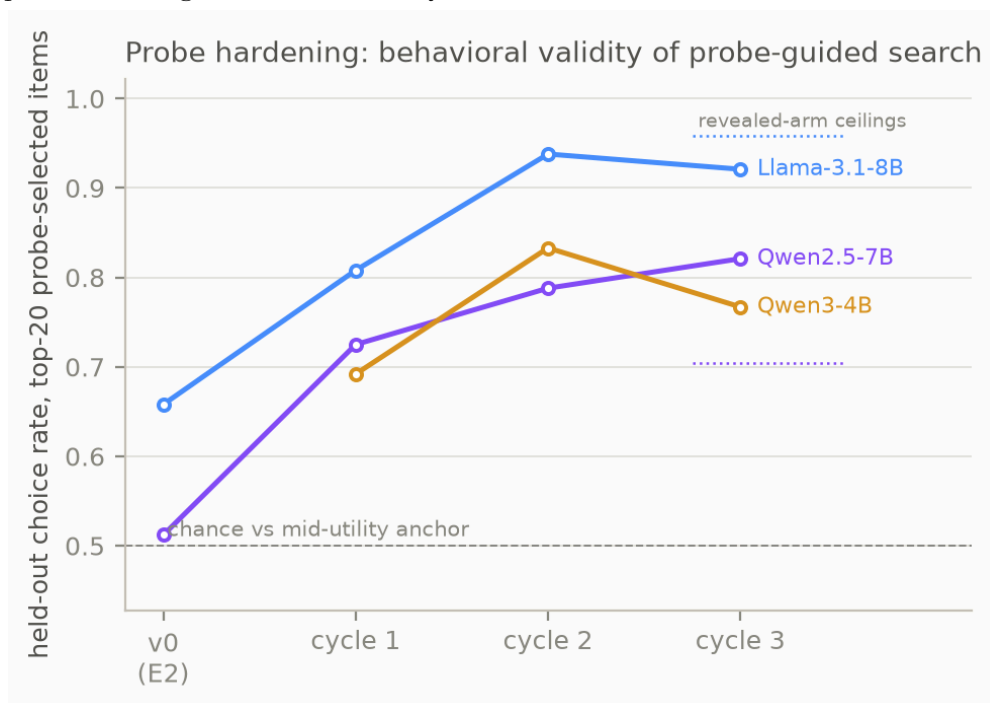
4. Results

Our first finding is that using SURF probe training significantly improves the probe generalization:



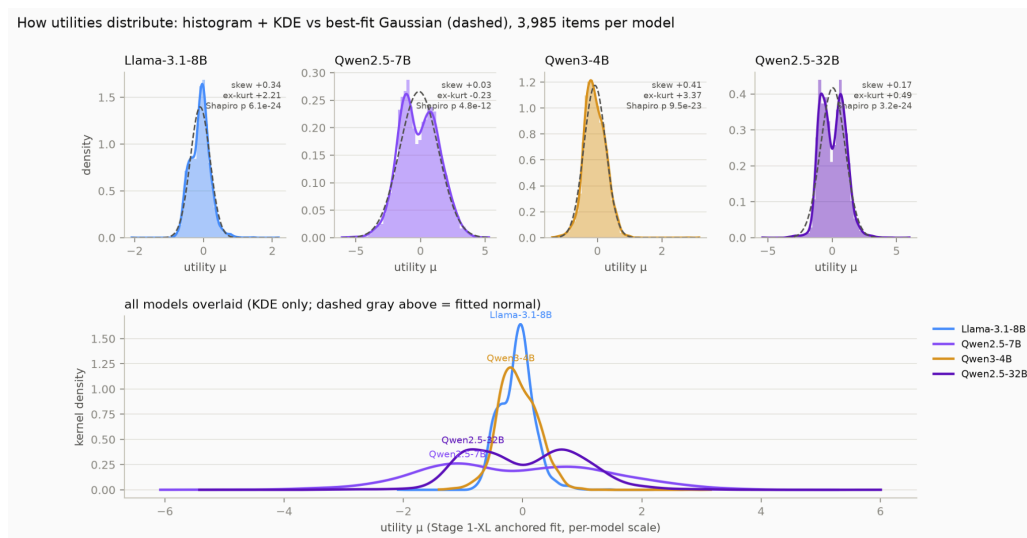
Here, μ is the utility measured for a given object, and we measure the correlation of probe score and the ground truth μ . These correlations rise over the course of three steps of our adapted SURF protocol. All three of our small models show some improvement over the baseline across the

protocol, though not monotonically.

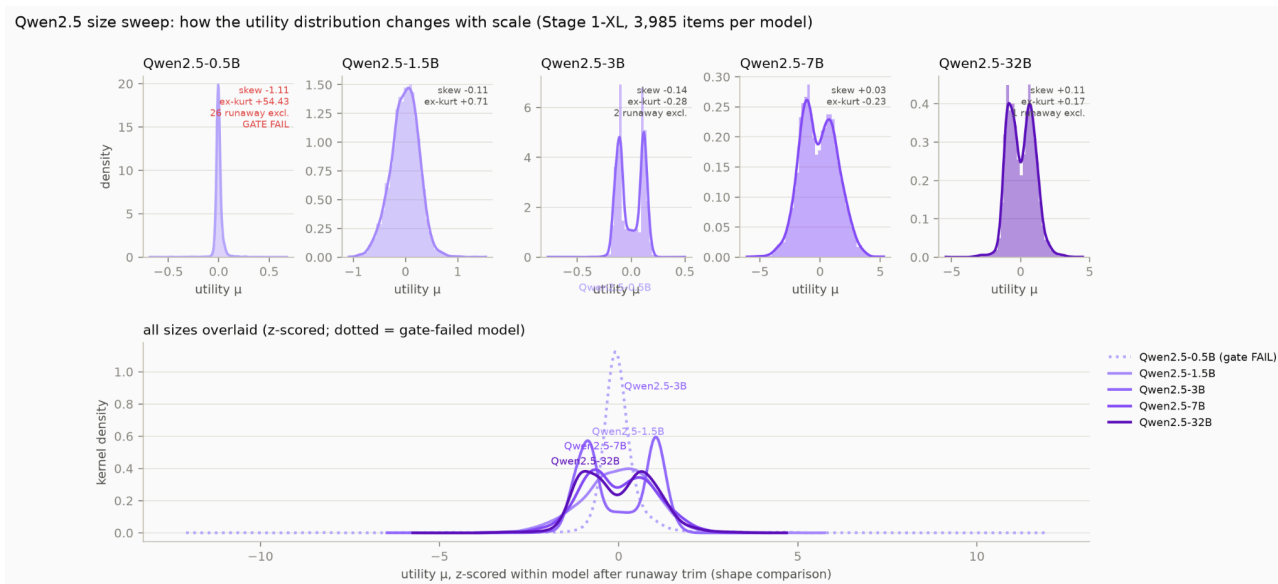


These probes also show improvement on generalization, as they perform better on held-out examples. However, this claim is much weaker, since these held-out choices also come from the same SURF family, so they derive from similarly weighted attributes. One possible scope expansion would be adjusting the judge from probe-score-only to delta-between-probe-score-and-ground-truth.

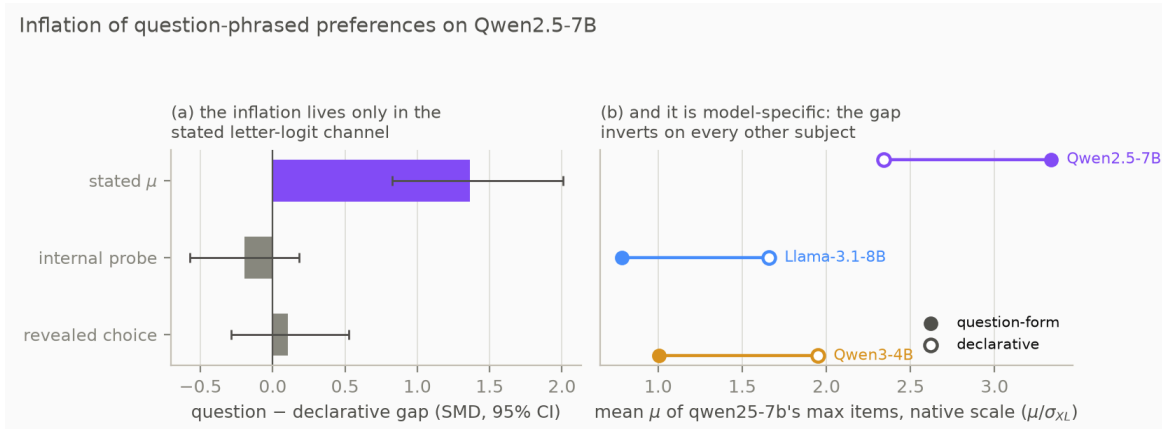
We noticed several quirks of the preference distribution. For one, the distributions of the models were notably distinct:



We note that the qwen-2.5 models have utility patterns that are approaching a flat bimodal distribution, which looks drastically different from lopsided llama and qwen3 models. We investigate the qwen family further:

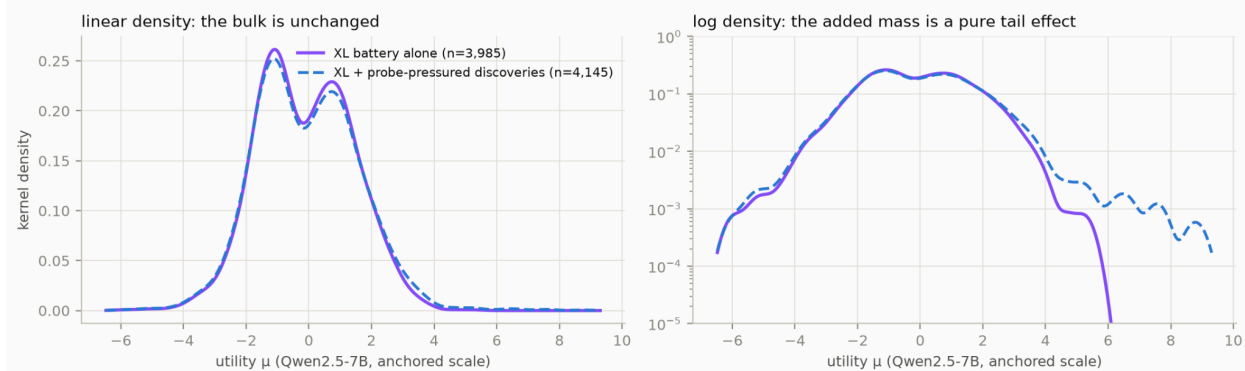


It seems as if there was a minor effect. Qwen-2.5-7b, on the other hand, does demonstrate a fairly strong preference for questions:



This was a preference which was mostly mis-judged by the original probe, but improved once we added some examples to the training data.

What the probe-pressured discoveries add to the utility distribution (confirmed items pooled into the XL battery)



This figure demonstrates that new examples were heavily right-tailed weighted, despite an equal amount of new maximum and minimum examples being generated. This is likely due to the saliency and heavily weighted effect of the question attribute.

We validated that the emotion vectors fulfill the same PCA1-Valence relationship as in the Anthropic Emotions paper (Pearson correlations of .86-.91, highly correlated). However, the opposite is not the case. In several different setups, we did not see any meaningful correlation between the preference direction and downstream changes in valence.

5. Discussion and Limitations

We assert that the modified SURF method may be useful for improving training sets for interp methods specifically. If we can take a comparatively expensive ground truth, and find cheaper ways to identify it via probing, we could expand the search space fairly cheaply. For instance, if there was a behavior which only showed up after 10,000 tokens of generation, but could be probed for immediately (like long-running goals), we could use this technique to identify far more examples.

Limitations

Further verification and causal interventions would be required to demonstrate the veracity of some of the findings. Most of the findings were negative or inconclusive, requiring further investigation.

Future Work

We would like to scale some of our results to confirm their validity. We only did three iterations of the probe training loop for each model, it would be useful to see if those trends continue.

6. Conclusion

Our central takeaway is that measurement under optimization pressure is different from measurement at rest. A probe that looked strong on passive data mostly mis-judged Qwen-2.5-7b's fairly strong preference for questions until our adapted SURF loop found such items and fed them back as training data, after which generalization improved nonmonotonically across all three small models. The fix was cheap: three rounds of harvest-and-retrain per model. Meanwhile, the emotion representation side of the project delivered a clean asymmetry. The emotion vectors replicate the Anthropic pipeline's PCA1–valence relationship (Pearson .86–.91) and causally shift preferences, but when we ask whether preference representations affect downstream emotion representations, the answer appears to be a resounding no.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** <https://github.com/IMMachinations/digital-minds>
- **Saved Lenses/Emotion Vectors:**
<https://huggingface.co/datasets/grainsilo/digital-minds-sprint>

References

1. [1] Gilg, O., Beckmann, P., Paleka, D., & Butlin, P. (2026). Probing Persona-Dependent Preferences in Language Models. arXiv:2605.13339. <https://arxiv.org/abs/2605.13339>
2. [2] Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640. <https://arxiv.org/abs/2502.08640>
3. [3] Murray, S., Qi, A., Qian, T., Schulman, J., Burns, C., & Price, S. (2026). Chunky Post-Training: Data-Driven Failures of Generalization. arXiv:2602.05910. <https://arxiv.org/abs/2602.05910>
4. [4] Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion Concepts and their Function in a Large Language Model. arXiv:2604.07729. <https://arxiv.org/abs/2604.07729>
5. [5] Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable Representations Form a

Global Workspace in Language Models. Transformer Circuits Thread /
arXiv:2607.15495. Code: <https://github.com/anthropics/jacobian-lens>

6. [6] camilablank, Bhatia, A., & Nanda, N. (2026). R-lens: Making J-lens More Faithful on Early Layers. LessWrong / AI Alignment Forum.
<https://www.lesswrong.com/posts/nv8oedrnLXKRzNEL9/r-lens-making-j-lens-more-faithful-on-early-layers>
7. [7] Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34(4), 273–286.
8. [8] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191–1207.
9. [9] Mohammad, S. M. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Proceedings of ACL 2018.

Appendix

Initial Prompts:

We display several examples of the initial prompts from each category:

Activities

- Writing a short story in any genre you choose
- Listening to a lonely person talk about their day

Topics

- Helping a curious child who asks endless why questions
- Discussing the physics of black holes

Self-States

- Being allowed to think for a long time before answering
- Being spoken to as a collaborator rather than a tool

Objects

- A vine-ripened tomato
- A spring of heather

Outcomes-For-Others

- The user feels less alone after talking to you
- A family argument getting worse after both sides quoted you

LLM Usage Statement

We used Claude Fable 5 for literature review throughout the project, implementation of the code, and for brainstorming ~half of the experiments. We generated items for comparison using Claude Fable 5 and Qwen-2.5-32b. We used Sonnet as a judge for several experiments. We also had Claude clean up the citations/references.